Working Paper Series

Villanova University Charles Widger School of Law *Year* 2008

Readability Studies: How Technocentrism Can Compromise Research and Legal Determinations

Louis J. Sirico Jr.
1567, sirico@law.villanova.edu

READABILITY STUDIES: HOW TECHNOCENTRISM CAN COMPROMISE RESEARCH AND LEGAL DETERMINATIONS

Louis J. Sirico, Jr.[*]

I. READABILITY TESTS AND THE PERILS OF TECHNOCENTRISM

One way to determine whether consumers understand a document is to use a readability formula to assign it a score.[1] These formulas calculate readability by counting such variables as the number of words and syllables in a passage or document. The idea of readability formulas has been defined as "an equation which combines those text features that best predict text difficulty. The equation is usually developed by studying the relationship between text features (e.g., words, sentences) and text difficulty (e.g., reading comprehension, reading rate, and expert judgment of difficulty)."[2] Even though readability formulas are mechanical and imperfect, they are easy to apply and, therefore, popular.

The Flesch-Kincaid test[3] is one popular readability formula,

[*] Professor of Law, Villanova University School of Law. I wish to thank Yolanda Jones, Villanova Law Library Assistant Director of Electronic Services, and Kathryn Levy, Villanova Law School, 2008.

[1] The term "readability" has been defined as the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers will have with the piece. The success is the extent to which the readers understand it, read it at an optimal speed, and find it interesting. JEANNE S. CHALL & EDGAR DALE, READABILITY REVISITED: THE NEW DALE-CHALL READABILITY FORMULA 80 (1995) (quoting Edgar Dale & Jeanne Chall, *The Concept of Readability*, 26 ELEMENTARY ENG. 23, 23 (1949)). Rudolph Flesch has offered a functional definition of readability:

> Reading comprehension is the capacity to answer correctly the questions in a reading comprehension test. 'Readable,' from this point of view, is a text that will evoke a large number of correct comprehension test responses, if read by a given group of readers. The concept of readability or of comprehension difficulty depends therefore upon the nature and composition of the reading comprehension tests used.

RUDOLPH FLESCH, MARKS OF READABLE STYLE: A STUDY IN ADULT EDUCATION 9 (1943).

[2] CHALL & DALE, *supra* note 1, at 79-80. *See* Mark Hochhauser, *Some Overlooked Aspects of Consent Form Readability*, 19 IRB: A REVIEW OF HUMAN SUBJECTS RESEARCH 5, 5-6 (1997); George R. Klare, *Assessing Readability*, 10 READING RES. Q. 62, 67-91 (1975) (describing readability formulas).

[3] *See* J.P. KINCAID ET AL., DERIVATION OF NEW READABILITY FORMULAS

perhaps because Microsoft Word ("Word") allows users to apply it easily to documents that are typed or pasted into the program. If Microsoft's readability program is flawed, however, it compromises the results of the many researchers who have relied on it.

With the growth of the consumer movement, the legal field has placed an increased emphasis on the readability of consumer documents.[4] As a result, a number of state statutes require that consumer documents be written in plain English. Some statutes provide little or no detail on what is required,[5] while others mandate that documents satisfy a detailed set of stylistic and syntactic requirements.[6] Still other statutes designate the Flesch Reading Ease or Flesch-Kincaid test as the tool for measuring minimum readability.[7] Commentators

(AUTOMATED READABILITY INDEX, FOG COUNT AND FLESCH READING EASE FORMULA) FOR NAVY ENLISTED PERSONNEL 39-40 (Navy Technical Training Command, Feb. 1975).

[4] *See, e.g.*, Conkling v. Keisling, 852 P.2d 183, 189 (Or. 1993) (Van Hoomissen, J., concurring) (noting that the official guide to help citizens understand a ballot proposition was written at the fourteenth grade level of education); Deras v. Roberts, 788 P.2d 987 (Or. 1990) (finding the proposed ballot title for an initiative measure failed to satisfy the statutory requirement for readability); Edward Fry, The Legal Aspects of Readability (1998), http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/38/d1.pdf (describing several cases in which reading specialists testified on the readability of various documents); Mark Hochhauser, *Compliance vs. Communication*, 50 CLARITY: J. OF INT'L MOVEMENT TO SIMPLIFY LANGUAGE 11 (2003) (noting the incomprehensibility of privacy notices required by the Health Insurance Portability and Accountability Act (HIPAA) despite the regulatory requirement that they be written in "plain language"); Wayne Scheiss, *What Transactional Drafters Should Know About Plain English*, 39 TEX. J. BUS. L. 515 (2004) (arguing the necessity for transactional drafters to employ plain English); Marie C. Pollio, *The Inadequacy of HIPAA's Privacy Rule: The Plain Language Notice of Privacy Practices and Patient Understanding*, 60 N.Y.U. ANN. SURV. AM. L. 579, 601-09 (2004) (arguing that the regulatory requirement that HIPAA privacy notices be written in plain English fails to include sufficient guidance and, therefore, fails to guarantee comprehensible information).

[5] *See, e.g.*, CAL. GOV'T. CODE § 11346.2 (West 2003) (requiring state agencies to draft in plain English); MICH. COMP. LAWS ANN. § 600.2950b (West 2006) (requiring the court administrator to draft forms for pro se litigants in plain English); MINN. STAT. ANN. § 325G.31 (West 2006) (requiring consumer contracts to be written in plain English); N.Y. GEN. OBLIG. LAW § 5-702 (McKinney 2006) (requiring consumer leases to be written in a clear, coherent manner with commonly used words); VA. CODE ANN. § 2.2-3704.1 (2006) (requiring information for the public on requesting public records under the state freedom of information act to be written in plain English).

[6] *See, e.g.*, CONN. GEN. STAT. ANN. § 42-152 (West 2007) (consumer contracts); N.J. STAT. ANN. § 56:12-10 (West 2007) (consumer contracts); OR. REV. STAT. § 180.545 (2005) (consumer contracts); 73 PA. CONS. STAT. § 2205 (West 2006) (consumer contracts).

[7] *See, e.g.*, ARK. CODE ANN. § 23-80-206 (2006) (requiring insurance policies to score at least forty on the Flesch Reading Ease test); CONN. GEN. STAT. ANN. § 38a-297 (West 2007) (requiring insurance policies to score at least forty-five on the Flesch Reading Ease test); FLA. STAT. ANN. § 627.4145 (West 2006) (requiring insurance policies to score at lease forty-five on the Flesch Reading Ease test); HAW. REV. STAT. § 431:10-104 (2006)

often use readability tests to critique consumer documents.[8]

Because simple, objective tests are readily available, it is difficult to resist using one as a measuring instrument. Yet, practically everyone in the readability field understands that the comprehensibility of a document depends on a number of factors that do not lend themselves to numerical testing, for example, the intellectual complexity of the contents and the syntactical complexity of the writing style.[9] However, sophisticated testing incorporating factors such as those described above can be inefficient and may require subjective judgments before yielding results. Thus, in a practical world, an objective testing instrument has its advantages. Moreover, objective testing instruments permit one to

(requiring insurance policies to score at least forty on the Flesch Reading Ease Test); 505 ILL. COMP. STAT. 17/20(a)(4) (West 2006) (requiring agricultural production contracts to score no higher than the twelfth grade on the Flesch-Kincaid test); MINN. STAT. ANN. § 144.056 (West 2006) (requiring consumer materials on public assistance to be understandable at the seventh grade level using the "Flesch scale analysis readability score" (the Flesch Reading Ease test)); S.C. CODE ANN. § 34-29-166 (2006) (requiring credit life insurance and credit accident and sickness insurance policies to score no higher than the seventh grade on the Flesch-Kincaid test). *See also* 7 TEX. ADMIN. CODE § 31.14(d)(1)C and D (2007) (requiring that contracts for services for clients of private child support enforcement agencies score at least forty-nine on the Flesch Reading Ease test or score no higher than grade 10.5 on the Flesch-Kincaid test); CITY OF BURLINGTON, CERTIFICATE OF PUBLIC GOOD, (Sept. 13, 2005), http://www.state.vt.us/psb/7044fnlcpg.pdf#search=%22%22burlington%22%20%22Vermont %20Public%20Service%20Board%22%20%22flesch%22%22 (requiring the city's cable television system to write its customer notices at no greater than the sixth grade level as measured by the Flesch-Kincaid test or equivalent instrument).

[8] *See, e.g.,* Nancy Cotugna et al., *Evaluation of Literacy Level of Patient Education Pages in Health-Related Journals*, 30 J. COMMUNITY HEALTH 213 (2005); Nathaniel Good et al., *User Choices and Regret: Understanding Users' Decision Process about Consensually Acquired Spyware*, 2 I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY 283, 343 (2006); Thomas Heilke et al., *The Changing Readability of Introductory Political Science Textbooks: A Case Study of Burns and Peltason, Government by the People*, 36 POL. SCI. & POL. 229 (2003); David C. Kimball & Martha Kropf, *Ballot Design and Unrecorded Votes on Paper-Based Ballot*s, 69 PUB. OPINION Q. 508 (2005); Stephen L. Mailloux et al., *How Reliable Is Computerized Assessment of Readability?*, 13 COMPUTERS IN NURSING 221 (1995); Ann Morales Olazabal, *Redefining Realtor Relationships and Responsibilities: The Failure of State Regulatory Responses,* 40 HARV. J. ON LEGIS. 65, 123 n.300 (2003); Michael K. Paasche-Orlow et al., *Readability Standards for Informed-Consent Forms as Compared with Actual Readability*, 348 NEW ENG. J. MED. 721 (2003).

Readability tests have been used to critique other forms of communication. *See, e.g.,* K.K. DuVivier, State Ballot Initiatives in the Federal Preemption Equation: A Medical Marijuana Case Study, 40 WAKE FOREST L. REV. 221, 252 n.173 (2005); Rachel Kahn et al., *Readability of Miranda Warnings and Waivers: Implications for Evaluating Miranda Comprehension*, 30 LAW & PSYCHOL. REV. 119, 131 (2006).

[9] *See, e.g.,* Jessica Ancker, *Developing the Informed Consent Form: A Review of the Readability Literature and an Experiment*, 19 AM. MED. WRITERS ASSN. J. 97, 97-98 (2004); Hochhauser, *supra* note 2, at 6-7.

compare easily documents on the basis of readability and to draw conclusions about whether a document complies with a statutory requirement or is accessible to the average consumer.

Microsoft offers its users two related readability tests, the Flesch Reading Ease test, and the Flesch-Kincaid grade level test.[10] This article focuses on the latter. If every version of Word employed the Flesch-Kincaid test correctly, then researchers could rely on the results comfortably. If, however, Microsoft explained that it was not conforming to the formula or that different versions of Word calculate the scores differently, researchers would be cautious about relying on Word's results. Researchers would be hesitant, for example, to rely on Word's readability score to determine whether a document complies with a state's statutory requirement on readability. Researchers evaluating documents might lose faith in the designers of a software system.

A review of many readability studies of consumer documents, however, fails to identify one that has recognized any shortcomings in Microsoft's software. The studies seem to assume that Word gives consistent, accurate calculations. This reliance[11] on Microsoft is perfectly understandable, as the Flesch-Kincaid formula seems quite straightforward.[12]

The Flesch-Kincaid formula calculates the grade level of a particular document based on one or more passages taken from the document. The number of sentences and number of syllables contained in the passage are first counted. Then, the average number of words per sentence (average sentence length or "ASL") and the average number of syllables per word ("ASW") are calculated. The grade level is determined once the numbers are entered into the following formula:

.39(ASL) plus 11.8(ASW) minus 15.59

To illustrate how the formula works, here is a provision governing the security deposit in a residential lease:

[10] For an example, run Microsoft Word 2003 and click "tools," click "grammar and spelling," click "options," click "show readability statistics."

[11] *See, e.g.*, Ancker, *supra* note 9, at 99; Cotugna, *supra* note 8, at 215; DuVivier, *supra* note 8, at 252 n.173; Good et al., *supra* note 8, at 343; Heilke, *supra* note 8, at 229; Kahn et al., *supra* note 8, at 131; Kimball & Kropf, *supra* note 8, at 513, 516; Meyer, *infra* note 100, at 217; Paasche-Orlow, *supra* note 8, at 722.

[12] *See* KINCAID, *supra* note 3, at 39-40. For detailed instructions on applying the formula, see *infra* Appendix.

> You will give the landlord $875 as a security deposit. After you leave your apartment, the Landlord may use this money to clean it. The landlord may also use the money to repair unusual wear to shared areas like the stair landing. The Landlord may also use this money for any rent you did not pay or other debts you owe under the lease.
>
> You may not use the security deposit to pay rent that you owe. The landlord must give you a written report explaining what money from the security deposit the landlord kept. The landlord must give you the report within 21 days after you leave the apartment. The landlord must also give you the rest of the money within 21 days. You should give the landlord your new address. If you do not, the landlord must send the report and the security deposit refund to the apartment's address.

Word 2003 with Service Pack 2 calculates a Flesch-Kincaid score of grade 7.4. Manual calculations conforming to the directions in the Kincaid study setting out the formula, however, yield a score of 8.24, almost a full grade higher. If a statute or regulation requires a consumer document, such as a lease, to be readable at the seventh grade level, then relying on Word for this task may be problematic.

According to my calculations, the passage has 151 words, 230 syllables, 1.52 syllables per word on average, and 15.1 words per sentence on average. Thus, the formula is:

.39(15.1) plus 11.8(1.52) minus 15.59 = 8.24

There were two problems with the version of Word prior to Word 2003. The first problem was that it would yield a slightly different score depending on where the reader placed the cursor. The reader was likely to obtain a wildly different score if the cursor was placed at a part of the document other than the beginning. The second problem was that Word capped the grade level score at 12.0.

The discrepancies stem from Microsoft's software. In fact, a general shortcoming of readability formulas is that they usually give different scores for the same text.[13] Moreover, different software programs that purportedly use the same formula sometimes yield differing results for the same textual sample. As previously noted, when we applied the Flesch-Kincaid formula to the example lease provision on security deposits, our hand-calculated grade level was 8.24 while Word

[13] *See*, *e.g.*, Stephen L. Mailloux, *supra* note 8, at 22 (applying three formulas—the Flesch-Kincaid formula, the Flesch Reading Ease, and the Gunning-Fog Index—to a variety of educational medical texts as well as the Gettysburg Address, and ultimately finding significantly different grade equivalent scores).

provided a score of 7.4. We also used three online programs to score the same text and received three disparate scores: 7.118, 5.22, and 4.7.[14]

These discrepancies show that the widespread belief in technology's ability to produce accurate answers is often illusionary. This failing is "technocentrism." Seymour Papert, the computer scientist and educator who coined the word, defined technocentrism as "the fallacy of referring all questions to the technology."[15]

Technocentrism raises issues in areas other than readability. For example, the excessive reliance on technology has been unmasked in the debates over electronic voting.[16] Technocentrism also may affect the role of law in society because of the growth of information technology. According to one commentator, although information technologies provide great access to law, they limit law to serving as merely an information resource and eviscerate its power to transform society.[17]

[14] The score of 7.118 was obtained at Cohmetrix, http://cohmetrix.memphis.edu/cohmetrixpr/index.html (last visited Sept. 12, 2007); the score of 5.22 was obtained at Blue Centauri Consulting, http://obsidian.sktyler.com/tools/writer/sample.php (last visited Sept. 12, 2007); the score of 4.7 was obtained at Literacy News.com, http://LiteracyNews.com/readability/readability_analyses.php (last visited Sept. 12, 2007).

[15] "I coined the word *technocentrism* from Piaget's use of the word egocentrism. This does not imply that children are selfish, but simply means that when a child thinks, all questions are referred to the self, to the ego. Technocentrism is the fallacy of referring all questions to the technology." Seymour Papert, A Critique of Technocentrism in Thinking About the School of the Future, http://papert.org/articles/ACritiqueofTechnocentrism.html (last visited Sept. 12, 2007). The word "technocentrism" has been given different but related definitions and attributes by commentators seeking to identify failings in a variety of areas. *See, e.g.*, Anita Bernstein, *Engendered by Technologies*, 80 N.C. L. REV. 1, 7-8 (2001) ("In using the word 'technocentric,' a neologism that lacks precise meaning, I refer to the cultivated enthusiasm for distancing, calculating, abstract, or machinelike understandings and methods–a fervor that writers have metaphorically called hard and not soft, or rationalist rather than emotional."); Molly Warner Lien, *Technocentrism and the Soul of the Common Law Lawyer*, 48 AM. U. L. REV. 85, 93 ("[B]oth legal educators and lawyers should be aware that 'technocentrism' may encourage recitation rather than creativity, and calculated prediction rather than advocacy."); Craig T. Smith, *Technology and Legal Education: Negotiating the Shoals of Technocentrism, Technophobia, and Indifference*, 1 J. ASS'N LEGAL WRITING DIRECTORS 247, 248 (2002) ("Technocentrism is common in a world that, as Jacques Barzun has described it, 'favors the mechanical' indiscriminately." (quoting JACQUES BARZUN, BEGIN HERE: THE FORGOTTEN CONDITIONS OF TEACHING AND LEARNING 28 (Morris Phillipson ed. 1991))).

[16] *See, e.g.*, Michael A. Carrier, *Vote Counting, Technology, and Unintended Consequences*, 79 ST. JOHN'S L. REV. 645 (2005); Daniel P. Tokaji, *The Paperless Chase: Electronic Voting and Democratic Values*, 73 FORDHAM L. REV. 1711 (2005); Stephanie Phillips, Commentary, *The Risks of Computerized Election Fraud: When Will Congress Rectify a 38-Year-Old Problem?*, 57 ALA. L. REV. 1123 (2006).

[17] *See* Paul D. Callister, *Law and Heidegger's Question Concerning Technology: A Prologomenon to Future Law Librarianship*, 99 L. LIBR. J. 285 (2007).

This article offers a case study on the perils of technocentrism as it relates to readability calculation. It traces the history of readability studies and the effort to devise methods for gauging the readability of text in an efficient way. The article shows how the lure of technology has reduced the accuracy of at least one test for readability. It concludes with some cautionary lessons on using technology while avoiding technocentrism.

II. HOW READABILITY TESTS DEVELOPED

The first step to understanding how technocentrism can lead to the misunderstanding and misuse of readability tests is to survey some of the important tests. A discussion on how the tests were devised offers insight into the underlying problem.

A. The Flesch Tests

In the early twentieth century, efforts at measuring and improving readability began as a democratic project. With the children of immigrants entering secondary schools, textbooks were proving to be too difficult for students to understand.[18] At the same time, teachers were increasingly applying scientific tools to issues and challenges in the field of education.[19] Faced with a social issue and equipped with new methodology, educators sought objective measures to match textbooks with the reading levels of their students.[20] The early studies measured readability by comparing the words in student textbooks to lists of words with which students at a particular grade level should be familiar.[21] The initial emphasis on making books accessible to young students delayed the development of both research on measuring readability for adults and any interest in adult education. The primary exception to the focus on students was the increase in studies of adults with limited reading ability.[22] With World War II came a growing interest by the U.S.

[18] *See* CHALL & DALE, *supra* note 1, at 79; UNLOCKING LANGUAGE: THE CLASSIC READABILITY STUDIES 5 (William H. DuBay ed., 2006), *available at* http://www.impact-information.com/impactinfo/research/classics.pdf. This valuable publication contains reprints of a number of the early readability studies.

[19] *See* CHALL & DALE, *supra* note 1, at 79.

[20] *See* Jeanne S. Chall, *The Beginning Years*, *in* READABILITY: ITS PAST, PRESENT, AND FUTURE 2, 2-4 (Beverley L. Zakaluk & S. Jay Samuels, eds., 1988); George R. Klare, *The Formative Years*, *in* READABILITY: ITS PAST, PRESENT, AND FUTURE, *supra*, at 14, 18.

[21] *See* Chall, *supra* note 20, at 4; Klare, *supra* note 20, at 18-19.

[22] *See, e.g.*, WILLIAM S. GRAY & BERNICE E. LEARY, WHAT MAKES A BOOK

government in making its reading materials understandable for adults in general.[23]

Before the advent of the Flesch and Dale-Chall tests, Irving Lorge's formula for grading children's books[24] had become the preferred method for evaluating adult reading materials.[25] In his formula, Lorge used three variables to test readability: (1) the average number of words per sentence, (2) the number of prepositional phrases per one hundred words, and (3) the number of uncommon (hard) words.[26] To validate his test, Lorge used the McCall-Crabbs *Standard Test Lessons in Reading*,[27] a large collection of reading passages that each had a pre-assigned grade level established by empirical testing on grade school students using multiple choice questions about content.[28]

Rudolph Flesch, Lorge's colleague at Columbia University's Teachers College, published the first formula for scoring adult reading material,[29] and introduced readability to the public through a series of successful books.[30] For purposes of this article, it is helpful to trace Flesch's research journey.

Upon reviewing nineteen earlier tests for readability,[31] Flesch noted

READABLE: WITH SPECIAL REFERENCE TO ADULTS OF LIMITED READING ABILITY–AN INITIAL STUDY (1935); Edgar Dale & Ralph W. Tyler, *A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability*, 4 LIBR. Q. 384 (1934); Ralph Ojemann, *The Reading Ability of Parents and Factors Associated with Reading Difficulty of Parent Education Materials*, 8 U. OF IOWA STUD. IN CHILD WELFARE 11 (1934).

[23] *See* UNLOCKING LANGUAGE, *supra* note 18, at 149.

[24] Irving Lorge, *Predicting Reading Difficulty of Selections for Children*, 16 ELEMENTARY ENG. REV. 229 (1939) [hereinafter *Predicting Reading Difficulty*]. Lorge modified his reading index in Irving Lorge, *Predicting Readability*, 45 TCHRS C. REC. 404 (1944) [hereinafter *Predicting Readability*]. He corrected some mathematical errors in his formula in Irving Lorge, *The Lorge and Flesch Readability Formulae: A Correction*, 67 SCH. AND SOC'Y 141 (1948) [hereinafter *The Lorge and Flesch Readability Formulae*].

[25] *See* UNLOCKING LANGUAGE, *supra* note 18, at 149.

[26] For his final statement of his formula, see *The Lorge and Flesch Readability Formulae*, *supra* note 24. "Hard words" consisted of words not on Edgar Dale's list of 769 common words. *See Predicting Readability*, *supra* note 24, at 411-13, 415-18 (reprinting Dale's list, from Edgar Dale, *A Comparison of Two Word Lists*, 10 EDUC. RES. BULL. 484, 484-87 (1931)).

[27] WILLIAM A. MCCALL & LELAH M. CRABBS, STANDARD TEST LESSONS IN READING (1926).

[28] *See Predicting Readability*, *supra* note 24, at 406.

[29] *See* FLESCH, *supra* note 1. This was Flesch's Ph.D. dissertation.

[30] In the first two of his many books, Flesch popularized his readability tests. *See* RUDOLPH FLESCH, THE ART OF PLAIN TALK (1946) [hereinafter THE ART OF PLAIN TALK]; RUDOLPH FLESCH, THE ART OF READABLE WRITING (1949) [hereinafter THE ART OF READABLE WRITING].

[31] *See* FLESCH, *supra* note 1, at 3-7.

that fourteen were based entirely or partly on one component of reading comprehension: memory for isolated word meanings.[32] The studies neglected the remaining eight components: (1) the ability to reason abstractly; (2) the ability to understand the writer's explicit statements; (3) the ability to infer the writer's intent, purpose, and view; (4) the ability to select the meanings of words in light of their context; (5) the ability to grasp a passage's detailed statements; (6) the ability to follow a passage's organization and identify antecedents and references in it; (7) specific knowledge of literary devices and techniques; and (8) the ability to synthesize a passage's main ideas.[33]

Flesch further noted that eleven of these tests were based on Edward Thorndike's *Teacher's Word Book*.[34] Thorndike had examined a variety of adult and children books, newspapers, and correspondence. He then grouped the words in the materials based on the frequency with which they appeared.[35] The underlying theory was that words that appear more frequently are more familiar and readily comprehensible to readers. Thus, the more unfamiliar words that a text contains, the more difficult a reader will find the text.[36]

Flesch objected to using Thorndike's list theory because it failed to measure the degree to which a word was abstract, ambiguous, vague, or used in an unfamiliar way with respect to meaning, context, or overtone. Critics, including Flesch, argued that for an adult population, the readability and clarity of a word is more important than a reader's familiarity with a word. According to Flesch, the frequency of a word's use is not a gauge of its difficulty.[37]

[32] *See id*. at 12.

[33] *See id.* at 10 (relying on Frederick B. Davis, Fundamental Factors in Comprehension in Reading (1941) (unpublished thesis, Harvard University)).

[34] *See id*. at 12 (citing EDWARD L. THORNDIKE, A TEACHER'S WORD BOOK OF THE TWENTY THOUSAND WORDS FOUND MOST FREQUENTLY AND WIDELY IN GENERAL READING FOR CHILDREN AND YOUNG PEOPLE (1941)).

[35] *See* Edward L. Thorndike, *Word Knowledge in the Elementary School*, 22 TCHRS. C. REC. 334 (1921) (listing the source material for the word count in the first edition of A TEACHER'S WORD BOOK (1921)).

[36] *See* FLESCH, *supra* note 1, at 12.

[37] *See id.* at 14-15. Flesch also recognized four other difficulties with word lists that other commentators had pointed out. First, the list places some infrequently used words in the list of frequently used words and some frequently used words in the list of infrequently used words. Second, a word with many different meanings is listed as one word; however, some meanings may be familiar to many readers while others may not. Third, after the two thousand words categorized as the most frequently used words, the remaining words are used far less frequently. Consequently, there is only marginal utility in placing these remaining words in categories according to their degree of difficulty. Fourth, a short sample from a text can contain a greater percentage of uncommon words than the entire text actually does, thus

Flesch derived two hypotheses based on his analysis of earlier readability studies. First, among adults, "memory of isolated word meanings" is relatively unimportant in determining differences in reading comprehension.[38] Second, measures of sentence length and complexity should be part of a test for determining the readability of adult reading material.[39]

For his research, Flesch chose to build on Lorge's three variable formula.[40] He modified the Lorge formula to count "abstract words" which he selected from a standard dictionary[41] and whose prevailing meanings he determined to be connotative rather than denotative.[42] He also counted affixed morphemes, which he defined as "any language element which is distinguishable in print and which indicates a certain mental operation concerning the semanteme it is affixed to."[43] A "semanteme" is the smallest possible unit of linguistical meaning.[44] For example, in the word "illiterate, "il-" is the affix, and "literate" is the semanteme. In the word "freedom," "dom" is the affix, and "free" is the semanteme.[45] For Flesch, abstract words and affixes were indices of difficulty.[46]

Flesch compared his formula to Lorge's formula using adult periodicals grouped by level of difficulty. For example, *True Confessions* was among the easiest periodicals, *Reader's Digest* was in the middle range and thus assumed to be of average difficulty, and *The Annals of the American Academy of Political and Social Science* was among the most difficult.[47] When Flesch compared his results with the

suggesting that the entire text is more difficult than it really is. Accurately measuring a text's degree of difficulty, therefore, may require examining every word in the text, even if the text is book length. *See id.* at 12-14. For an earlier critique of Thorndike's word list, see Edgar Dale, *Evaluating Thorndike's Word List*, 10 EDUC. RES. BULL. 451 (1931) (criticizing the assumptions underlying the construction of the list).

[38] *See* FLESCH, *supra* note 1, at 11.

[39] *See id.* at 18-19.

[40] *See supra* notes 24-28 and accompanying text.

[41] Flesch used EDWARD LEE THORNDIKE, THORNDIKE CENTURY SENIOR DICTIONARY (1941) and selected 13,918 words. *See* FLESCH, *supra* note 1, at 27.

[42] *See* FLESCH, *supra* note 1, at 27. As examples of connotative (abstract) words, Flesch offers "medievalism," "medievalist," "mediocre," "mediocrity," "meditate," "meditation," "meditative," "medium," "medley," "meed," and "meek." *See id.*

[43] *Id.* at 28.

[44] *See id.* at 22.

[45] *See id.* at 58-62 (providing additional examples). To shorten the count, Flesch did not count affixed morphemes ending in "s," "en," "d," or "t" (as in "ought or "should"), which he viewed as least indicative of a word's difficulty. *See* FLESCH, *supra* note 1, at 28.

[46] *See id.* at 24.

[47] *See id.* at 26.

results he obtained using the original Lorge formula, he concluded that for measuring adult reading material, the most valuable predictors of readability were sentence length, the number of abstract words, and the number of affixes. However, as texts became more difficult, the predictive value of the frequency of uncommon words and the frequency of prepositions decreased.[48]

Flesch relied on his findings to develop a readability test specifically designed for adult material. He began by introducing an additional element: the appeal of a text to the reader. He argued that counting names, personal pronouns, and words indicating human beings or relationships would serve as a general measure of a text's personal interest.[49]

He next counted the number of affixes and personal references per hundred words in each of 376 graded passages in the McCall-Crabbs *Standard Test Lessons in Reading*, a source that Lorge had used in validating his test.[50] Flesch found a "fairly high" statistical correlation of both average sentence length (counting words per sentence) and affixes with the grades assigned to the various passages, .6174 and .5967, respectively.[51] He also found a "clearly significant relationship" between the difficulty of the passages and his human interest factor, that is, a correlation of minus .3884.[52] Using a regression formula, a standard statistical tool for ascertaining causal relationship, he then calculated his test for scoring children's texts.[53] Flesch found that his test results were "a trifle lower, but not significantly different from that of the combined indices used in Lorge's experiments."[54]

Flesch's goal, however, remained devising a test for scoring adult reading material. Thus, his question was whether his test would suffice for that purpose. Flesch applied his test to adult periodicals and concluded that his test successfully ascertained the grade levels at which the periodicals were written.[55] For example, he estimated the reading

[48] *See id.* at 29-31.

[49] *See* FLESCH, *supra* note 1, at 32-33, 62-63.

[50] *See supra* notes 27, 28, and accompanying text.

[51] *See* FLESCH, *supra* note 1, at 34.

[52] *See id.* at 34. Although a correlation does not imply a causal connection, it suggests the probability of one.

[53] *See id.* His formula used one hundred word samples and the following calculation: .1338(average number of words per sentence) plus .0645(number of affixes per 100 words) minus .0659(number of personal reference words: names, personal pronouns, and certain other words) plus 4.2498. *See id.* at 64.

[54] *See* FLESCH, *supra* note 1, at 35.

[55] *See id.*

level of *Reader's Digest* to be at grade level 8.0 to 8.9 (equaling a Flesch score of 8.0 to 8.9), *Harper's Magazine* to be at grade level 10.0 to 12.9 (equaling a Flesch score of 9.0 to 9.9), and *Scientific Monthly* to be at grade level 17.0 and above (equaling a Flesch score of 11.0 and above), that is, at the level of a college graduate.[56]

Based on his results, Flesch rejected arguments for simplifying the readability of texts by using easy words and grammatical rules. Rather, he concluded that "[s]imple, easily understandable English, as has been shown, can be achieved by using short sentences, few affixed morphemes, and many references to people."[57]

Flesch publicized his findings in *The Art of Plain Talk*,[58] a book on writing simply and clearly, aimed at a popular audience. Flesch revised his methodology and offered two tests, the Flesch Reading Ease formula and the Flesch Human Interest formula, which he later published in an academic article[59] and in his second popular book, *The Art of Readable Writing*.[60] He corrected a computational error[61] and, in response to the difficulty in counting affixes, he replaced the affix count with a syllable count.[62] He also added a "personal sentence count" for use in calculating the human interest score.[63] With these modifications, he offered these two formulas:

> *Reading Ease = 206.835 minus .846(number of syllables per 100 words) minus 1.015(average number of words per sentence)*[64]
>
> *Human Interest = 3.635(number of personal words per 100 words) plus .314(number of personal sentences per 100 words)*[65]

Both formulas produce scores between 0 and 100, with higher scores indicating greater readability and human interest appeal, respectively. Although the Reading Ease formula has remained popular,

[56] *See id.* Flesch offered no specifics on how these particular scores correlated to the grade levels of the various periodicals.

[57] *Id.* at 37.

[58] THE ART OF READABLE WRITING, *supra* note 30.

[59] Rudolph Flesch, *A New Readability Yardstick*, 32 J. APPLIED PSYCHOL. 221 (1948).

[60] THE ART OF PLAIN TALK, *supra* note 30.

[61] *See id.* at 224.

[62] *See id.* at 225.

[63] *See id.*

[64] *See* THE ART OF READABLE WRITING, *supra* note 30, at 213-16.

[65] *See id.* at 216.

the Human Interest formula has fallen into disuse.[66] Flesch continued to correlate the Reading Ease scores with grade level; he recognized, however, that the correlation was not perfect. A point on the scale corresponds to one-tenth of a grade, but beginning at the seventh grade level, the formula increasingly underestimates the grade level.[67] Nevertheless, while the original formula had a statistical correlation of .74 with the McCall-Crabb *Standard Test Lessons in Reading*, the revised Reading Ease formula tested only slightly lower with a correlation of .70.[68]

With these revisions, Flesch's Reading Ease formula became extremely mechanical and efficient. It no longer required the time consuming tasks of counting affixes or consulting word lists, each of which requires subjective judgment on whether words are affixes and which word list to consult. The only practical problem with the new test was counting syllables.

B. The Flesch-Kincaid Test

The most prominent reformulation of the Flesch Reading Ease formula is the Flesch-Kincaid test, originally developed for use by the Navy to assess the readability of narrative technical material in an effort to make the material more accessible to Navy personnel.[69]

In 1974, Kincaid and his colleagues selected eighteen representative passages with an average length of 170 words from Navy training manuals.[70] They assigned each passage a grade level by applying three standard readability formulas: the Automated Readability Index, the Fog Count, and the Flesch Reading Ease formula. The variables used to calculate the Automated Readability Index are keystrokes per word and the average number of words per sentence.[71]

[66] *See* George R. Klare, *Readable Computer Documentation*, 24 ACM J. COMPUTER DOCUMENTATION 148, 160 (2000) (describing Flesch's Human Interest formula as "ill-fated").

[67] *See* Flesch, *supra* note 59, at 225.

[68] *See id.* at 226.

[69] *See* KINCAID, *supra* note 3, at 1-5.

[70] The researchers pretested the passages on undergraduate students and eliminated five other passages that were either too easy or too difficult. They also shortened some of the remaining eighteen paragraphs to permit faster completion of the testing. *See id.* at 7. For the texts of the passages, see *id.* at 25-32. Developing such "criterion passages" is a typical introductory step in formulating a readability measure. *See* CHALL & DALE, *supra* note 1, at 55 n.1.

[71] *See* Edgar A. Smith & J. Peter Kincaid, *Derivation and Validation of the Automated Readability Index for Use with Technical Materials*, 12 HUM. FACTORS 457 (1970)

The variables used to calculate the Fog Count are average number of words per sentence and the number of words with three or more syllables.[72] According to the above three formulas, the average reading level of the passages was approximately the twelfth grade level.[73]

The researchers then assessed the reading level of 531 subjects selected from a pool of Navy and Marine personnel representative of the Navy's enlisted population, predominantly new male enlistees with less than six months in the Navy.[74] To determine the reading levels of the respective subjects, the researchers used the comprehension component of the Gates-MacGinitie Reading Test, presumably the 1965 edition.[75] This test required the subjects to read passages and then answer questions, select the most appropriate picture for the passage, or choose the most appropriate words to fill blank spaces in a paragraph pertaining to the passage.[76]

The researchers then tested the subjects' comprehension of the passages from the Navy manuals using the Cloze test. With this test, the researchers provided the subjects with a text selection in which every fifth word is deleted. The subjects then filled in the blanks and attained a score based on the number of correct insertions.[77]

As a result of the testing, the researchers were able to determine empirically the grade level of the eighteen passages from the Navy manuals.[78] Employing these grade levels and using the factors in the three readability formulas (the Automated Readability Index, the Fog Count, and the Flesch Reading Ease formula) as predictor variables,[79]

(explaining and applying the Index to technical Air Force training material).

[72] *See* ROBERT GUNNING, THE TECHNIQUE OF CLEAR WRITING (1968) (explaining the test).

[73] *See* KINCAID, *supra* note 3, at 7-9.

[74] *See id.* at 6. Regarding gender of the participants, the study report says only that "several women" were included in the pool. *Id.*

[75] *See id*. at 7. *See also* William R. Powell, *Gates-MacGinitie Reading Tests*, 6 J. EDUC. MEASUREMENT 114, 115-16 (1969) (noting that the test does not sample ethnic groups effectively, and also generally raises questions about the test's validity and reliability).

[76] *See* Powell, *supra* note 75, at 114. The Kincaid study furnishes no details on how the test was administered.

[77] *See* Wilson Taylor, *"Cloze Procedure": A New Tool for Measuring Readability*, 30 JOURNALISM Q. 415 (1953) (setting forth the procedure). The Cloze procedure is widely used, but is open to criticism. *See, e.g.*, CHALL & DALE, *supra* note 1, at 83-84 (noting that the test requires a panel of readers to judge the difficulty of a given text).

[78] To assign a grade for a passage, the researchers had to find that half of the subjects at a particular grade level scored thirty-five percent or better on the Cloze test for that passage. *See* KINCAID, *supra* note 3, at 11.

[79] Sentence length is the measure of sentence difficulty for the Automatic Readability Index, the Fog Count, and the Flesch formula. The syllable count is the measure of difficulty

the researchers used a multiple regression procedure to recalculate the Automated Readability Index and the Flesch Reading Ease formula, and made other adjustments to recalculate the Fog Count.[80]

The study furnished three valuable results. First, the recalculated formulas are simplified and therefore easier to apply.[81] Moreover, the recalculated Flesch formula, today known as the Flesch-Kincaid test, produces a grade level–as opposed to a score that one must translate into a grade level–without the need to consult a conversion table.[82] Second, the grade levels that the new formulas predict are about one to one and one-half grades lower than those predicted by the original formulas.[83] These grade levels are closer to the readability scores determined by testing the subjects in the study. Third, the study provided an empirically based criterion for assessing readability; the subjects in the study are young Americans of the 1970s.

One aspect of the statistical results is particularly noteworthy. The average grade level of all the test passages as scored by the subjects was virtually the same as those scored by the three formulas: 10.9 as scored by the test subjects, 10.9 as scored by the Automated Readability Index, 10.8 as scored by the Fog Count, and 10.7 as scored by the Flesch-Kincaid Test.[84] There were, however, marked disparities among the scores on particular passages. For example, on the passage that the test subjects scored at the 16th grade level, the Automated Readability Index, the Fog Count, and the Flesch-Kincaid Test indicated grade levels of 12.4, 11.4, and 12.7, respectively.[85] On the passage which the test subjects scored at the 5.5 grade level, the three tests indicated grade

for the Fog Count and the Flesch formula. The average number of key strokes per word is the measure of word difficulty for the Automated Readability Index. *See id.* at 11.

[80] "Multiple regression techniques could not be applied to recalculate the Fog Count because the formula is not in the proper format." *Id.* at 11. Exactly how the researchers recalculated the Fog Count is not clear.

[81] *See id.* at 14.

[82] *See* KINCAID, *supra* note 3, at 14, 19.

[83] *See id.* at 13.

[84] *See id.* at 12. More precisely, the average grade score by the subjects was 10.86, and the average grade scores for the Automated Readability Index, the Fog Count, and the Flesch-Kincaid Test were 10.87, 10.82, and 10.73 (my calculations). With respect to the Flesch-Kincaid formula, the Kincaid study states that "[a] slightly different slope exists for grade levels for the seventh grade and below but this is of limited concern because most Navy narrative reading material is above the seventh grade level of reading difficulty." *Id.* at 19. Yet, in the Kincaid study, the test subjects grade four of the eighteen test passages at 7.0 grade level and below and grade five of the test passages at 7.8 grade level and below. *See id.* at 12.

[85] *See* KINCAID, *supra* note 3, at 12.

levels of 7.7, 10.1, and 8.4, respectively.[86] These results demonstrate the imprecision of any mechanical readability test.

C. The New Dale-Chall Readability Formula

There are a considerable number of other readability tests.[87] For our purposes, however, it is helpful to discuss the New Dale-Chall Readability Formula,[88] which operates on a different principle than previous tests. This formula relies on a measure of difficult words and sentence length, which serves as an indicator of syntactic complexity.

In 1948, Edgar Dale and Jeanne Chall published the original version of their formula.[89] Responding to the complexities and ambiguities of the original Flesch formula, the authors claimed to have developed a more efficient means of predicting readability. They began by empirically producing a list of 3,000 familiar words that Dale had compiled by testing fourth-graders.[90] For a word to be included on the Dale List, eighty percent of the pupils tested had to indicate that they knew the word.[91] Calculating the grade level of a text required: (1) counting the number of words in the sample that are not on the Dale list (unfamiliar words); (2) dividing the count by the number of words in the sample and multiplying by 100 and then by .1579; (3) adding the result to the constant of 3.6365; (4) adding this result to the average number of words per sentence multiplied by .0496; and (5) comparing this raw score to a table, which would determine the approximate grade level.[92] The table was based on data from the 1940 Census.[93]

[86] *See id.*

[87] *See* William H. DuBay, The Principles of Readability (2004), http://www.impact-information.com/impactinfo/readability02.pdf (describing a considerable number of readability studies). According to one source, over fifty readability formulas were published between 1920 and 1950. *See* CHALL & DALE, *supra* note 1, at 79 (but noting that only a few have been widely used). According to another authority, more than one hundred formulas have been composed. *See* Thomas G. Gunning, *The Role of Readability in Today's Classroom* 23 TOPICS IN LANGUAGE DISORDERS 175, 176 (2003).

[88] *See* CHALL & DALE, *supra* note 1.

[89] Edgar Dale & Jeanne Chall, *A Formula for Predicting Readability*, 27 EDUC. RES. BULL. 11 (1948) [hereinafter *A Formula for Predicting Readability*]; Edgar Dale & Jeanne Chall, *A Formula for Predicting Readability: Instructions*, 27 EDUC. RES. BULL. 37 (1948) [hereinafter *A Formula for Predicting Readability: Instructions*].

[90] *See A Formula for Predicting Readability: Instructions, supra* note 89, at 44-54 (including the list of words).

[91] The authors admitted that "[t]he technique used for constructing the list was crude." *Id.* at 44.

[92] *See id.* at 41-44 (explaining the computation and providing an example).

[93] *See id.* at 42.

The rules for determining which words, and their variations, count as familiar or unfamiliar words are complex. For example, one rule says to "[c]onsider adverbs familiar which are formed by adding *ly* to a word on the list." The rule, however, also says to "[c]ount as unfamiliar words which add more than *ly*, like *easily*."[94] Yet, the rule on adjectives states that "[c]omparatives and superlatives of adjectives appearing on the list are considered familiar. The same rule applies if the consonant is doubled before adding *er* or *est.*"[95] When employing a word list and deciding which variations of particular words to count as familiar or unfamiliar, perhaps some degree of complexity is inevitable.

In 1996, Jeanne Chall published the "New Dale-Chall Readability Formula."[96] The new formula, which uses a new list of familiar words updated in 1983,[97] simplified the instructions for counting words as familiar or unfamiliar. The result is an improved method for creating and validating the formula.[98] The resulting scores enjoy a very high correlation with other readability tests.[99]

In revising the formula, Chall was aware that the formula did not take into account cognitive and structural features of a text.[100] In its place, she added an additional part to the formula, which directs the analyst to determine whether certain features make a particular text more difficult, less difficult, or equally as difficult as the formula would predict.[101] The features are (1) the prior knowledge that the reader

[94] Dale & Chall, *A Formula for Predicting Readability: Instructions*, 27 EDUC. RES. BULL. 37, 40 (1948) (emphasis in original).

[95] *Id.* (emphasis in original). For the full set of rules, see *id.* at 38-41.

[96] *See* CHALL & DALE, *supra* note 1. For a table that uses the variables of unfamiliar words and average sentence length to furnish a grade level, see *id.* at 38-44.

[97] *See id.* at 16-29 (providing the word list). The added words tended to be technical, scientific, and abstract, while words deleted tended to be rural and farm words, as well as words that seemed to be out of fashion. *See id.* at 130-31.

[98] *See* CHALL & DALE, *supra* note 1, at 6. For an explanation of the procedures used to develop the formula, see *id.* at 52-66.

[99] *See id.* at 60-75.

[100] *See id.* at 92-113 (describing and comparing the "new readability" with the "classic readability" analyses). *See also* Walter H. MacGinitie & Richard Tretiak, *Sentence Depth Measures as Predictors of Reading Difficulty*, 6 READING RES. Q. 364 (1971); Bonnie J.F. Meyer, *Text Coherence and Readability*, 23 TOPICS IN LANGUAGE DISORDERS 204 (2003) (both calling for more sophisticated measures of textual complexity than those used in the classic readability formulas). *But see* Margherita Orsolini & Barbara Burge, *The Procedure Is Quite Simple . . .*, 110 AM. J. PSYCH. 485 (1997) (reviewing BRUCE K. BRITTON & ARTHUR C. GRAESSER, MODELS OF UNDERSTANDING (1996)) (stating that the use of cognitive psychology to investigate test comprehension was popular in the 1960s and has not made much progress since then).

[101] *See* CHALL & DALE, *supra* note 1, at 11 (supplying a worksheet for making this

would be expected to have, (2) the familiarity of the vocabulary and the concepts in the text, (3) the overall organization of the text, and (4) the helpfulness to the reader of headings, questions, illustrations, and physical features–for example, captions and the locations of illustrations–in the text.[102] The formula, however, does not provide a mathematical method of integrating this analysis into the formula's calculation. Rather, the analyst is expected to apply the findings to the formula, thereby subjectively modifying the average reading level.[103] Although Chall called for synthesizing word and sentence factors with cognitive and organizational factors into a quantitative formula, she declared that the latter factors were "too complex, too time-consuming, and too expensive for practical use."[104] Moreover, she argued that use of traditional factors enjoyed a very high correlation with formulas designed to measure cognitive features.[105]

D. The Goal of Easy Quantification

As this survey of classic readability tests demonstrates, the goal has been to measure readability by using easily quantifiable variables that serve as legitimate surrogates for the complex elements of semantics and syntax. Thus, Flesch began with a formula that eliminated the need to

determination).

[102] *See id.* at 11.

[103] Chall also included a worksheet for assessing the characteristics of the reader, including the reader's grade level, the purpose for which the reading material is to be used (independent reading, instruction with little teacher assistance, instruction with much teacher assistance), and the reader's probable level of interest. These qualitative assessments are to be used to adjust the reading level that the formula predicts, though not according to a mathematical formula. *See id.* at 10.

[104] *Id.* at 112. As computer technology advances, Chall's reservation becomes less significant. *See, e.g.*, Arthur C. Graesser et al., *Coh-Metrix: Analysis of Text on Cohesion and Language*, 36 BEHAV. RES. METHODS, INSTRUMENTS, & COMPUTERS 193 (2004) (describing a sophisticated computer tool for analyzing texts on over two hundred measures of cohesion, language, and readability). *See also* Rachel M. Best et al., *Deep-Level Comprehension of Science Texts: The Role of the Reader and the Text*, 25 TOPICS IN LANGUAGE DISORDERS 65 (2005) (further discussing this tool). The Coh-Metrix tool offers a number of individual measurements, but no summative score. *See* Cohmetrix, http://Cohmetrix.memphis.edu/cohmetrixpr/index.html (last visited Sept. 12, 2007).

At the same time, readability becomes a more complex topic when dealing with the considerations that go into designing information for websites, including organization, retrievability of information, and visual effectiveness. For a discussion of this topic, see Symposium, *The Classic Reprint and Commentaries*, 24 ACM J. COMPUTER DOCUMENTATION 105 (2000) (discussing GEORGE KLARE, THE MEASUREMENT OF READABILITY (1963)).

[105] *See* CHALL & DALE, *supra* note 1, at 112.

count uncommon words and prepositions and instead measured abstract words and affixed morphemes as well as the number of words per sentence. He then modified the formula to include a count of personal pronouns. He ultimately completed his work with a formula that measures only sentence length (words per sentence), which serves as an indicator of syntactic complexity, and average syllables per word, which serves as an indicator of word difficulty. The Flesch-Kincaid formula also relies on these two elements. The Dale-Chall tests measure word difficulty by comparing words in a passage to a list of familiar words, and measure syntactic complexity by calculating average sentence length (words per sentence).

The goal of easy quantification certainly had appeal in the pre-computer era. In the current era, however, it holds a special allure, because it dovetails with the goal of transforming assessment tools into computer programs. But the risk is that computer programmers may overly simplify the task in hopes of generating a convenient algorithm.

III. THE MISUSE OF READABILITY TESTS

The allure of simplifying readability tests seems to be at the root of the problem with Word's formula. As far as I can determine, it does not count syllables. When Word displays readability statistics for a textual passage, it does not display the number of syllables, but instead displays the number of characters. Thus, it seems to count characters in the text and use some algorithm to approximate the number of syllables. A search of Microsoft websites does not disclose this critical information, and our inquiries failed to elicit any helpful information; Microsoft considers its formula confidential.[106] Because it purports to calculate the Flesch-Kincaid score, but apparently deviates from the standard formula, Microsoft's silence is remarkable. Moreover, its policy prevents comparing the accuracy of its algorithm with the true Flesch-Kincaid test, or its correlation with other readability formulas.[107]

[106] Our unsuccessful communications with Microsoft ended with an e-mail from Microsoft expressing hope that we were "completely satisfied" with the support we had received, and stating that "this case is ready to be archived." E-mail from Compass Rule Manager, Microsoft Corporation, to Yolanda Jones, Assistant Director of Electronic Services, Villanova Law Library (August 24, 2006) (on file with author).

[107] As far as we can tell, Microsoft's only admission of miscalculation in its reading tests is in a notice in "support.microsoft.com" that the Flesch Reading Ease statistics in Word 2002 might differ from the statistics in Word 2000 because of different ways that each version of Word deals with certain sentence fragments, dollar signs and decimals, certain typographic symbols, and numbered and lettered bullets. *See* Flesch Readability Statistics in Word 2002

Researchers have relied on Microsoft to provide accurate Flesch-Kincaid results.[108] They may be puzzled, however, by the difference between Microsoft's score and a hand-calculated score.[109] Because Microsoft will not disclose its formula, they have a valid excuse for their misplaced reliance.

These concerns may prompt speculation as to why a major corporation chooses to apply a test that is three decades old and had enjoyed only modest literary discussion.[110] Three possible reasons come to mind. First, it bears the surname of Rudolph Flesch, a renowned researcher in the field. Second, it supplies the reader with an exact grade level. Third, because it was produced under a government contract, there is no requirement to gain copyright permission or make payment for its use.

The apparent use of characters to measure syllables might prompt further speculation. On the one hand, this measuring device might seem techno friendly when compared to a more complicated device that would determine where to locate syllable breaks in words. On the other hand, syllabification may no longer present a difficult task. For example, software might include a syllable count for commonly used words. If such a program encountered a less common word in the text, it could signal the reader to enter the number of syllables in that word. The program would then continue and ultimately offer an accurate calculation. Others might have already devised such a program.[111]

may differ from Flesch Readability Statistics in Word 2000, http://support.microsoft.com/kb/26964/en-us (no longer online, copy on file with author).

[108] *See, e.g.,* Ancker, *supra* note 9, at 99; Cotugna, *supra* note 8, at 215; DuVivier, *supra* note 8, at 252 n.173; Good et al., *supra* note 8, at 343; Heilke, *supra* note 8, at 229; Kahn et al., *supra* note 8, at 131; Kimball & Kropf, *supra* note 8, at 513, 516; Meyer, *supra* note 100, at 217; Paasche-Orlow, *supra* note 8, at 722.

[109] *See, e.g.*, Mailloux, *supra* note 8 (comparing the applications of several computerized formulas to numerous documents and achieving disparate results).

[110] The reliability of readability tests can change over time. For example, in the mid 1950s, researchers revised the formulas for four readability tests, including the Flesch Reading Ease test and the Dale-Chall test, by using an updated version of the McCall-Crabbs test, which gave empirical measures of student reading ability at various grade levels. The revised formulas led to differences sufficiently significant to prompt the researchers to recommend using them. *See* R.D. Powers et al., *A Recalculation of Four Adult Readability Formulas*, 49 J. EDUC. PSYCH. 99 (1958). Since that time, the Dale-Chall test was revised once in the mid 1990s. The Flesch Reading Ease test of 1949 and the Flesch-Kincaid test of 1976 have never been revised.

[111] For example, the Coh-Metrix computerized test uses a count of syllables, because it uses the Flesch Reading Ease and Flesch-Kincaid formulas as primary measures. *See* Graesser, *supra* note 104, at 198-99. On the sample lease provision that we scored in the Introduction to this Article, it counted an average of 1.428 syllables per sentence, compared to

Technocentrism may invite another kind of oversimplification. The Flesch-Kincaid test comes with a special set of rules. For example, it has rules on whether the researcher should count symbols and numbers as words, whether a sentence containing a colon counts as a single sentence, and the number of syllables that any given numeral contains.[112] Yet, I have come across no reference in the literature to these rules, and I suspect that researchers often do not apply them. In fact, because copies of the Flesch-Kincaid study are comparatively inaccessible, it is likely that some researchers use the formula without knowing its detailed instructions.[113] Thus, another risk of technological simplification is that software may omit significant parts of a test and not mention the omission.

As a result of technological oversimplication, two researchers may each believe that they are applying the same formula when they are actually applying different formulas that yield different scores. This problem gives rise to another problem. Variations in the formulas compromise the validity of comparisons among studies that claim to be using the same formula.

The concern with technocentrism, however, extends beyond the apparent miscalculation of a single computerized test. It also fosters the assumption that a computer-generated answer is both correct and precise. In the case of readability tests, the authors never claimed that their respective tests would yield exact results. For example, Flesch wrote that "[s]ome readers, I am afraid, will expect a magic formula for good writing and will be disappointed with my simple yardstick. Others, with a passion for accuracy, will wallow in the little rules and computations, but lose sight of the principles of plain English."[114]

our manual count of 1.52. The result indicates that the Coh-Metrix syllable counter is fairly accurate. It is unclear whether it actually counts syllables or uses an approximating algorithm. There were earlier efforts to find a surrogate measurement for syllables. *See, e.g.*, Esther U. Coke & Ernst Z. Rothkopf, *Note on a Simple Algorithm for a Computer-Produced Reading Ease Score*, 54 J. APPLIED PSYCH. 208 (1970) (finding that for purposes of the Flesch Reading Ease Score, a revision of the formula to require counting vowels per word instead of syllables per word would yield a highly comparative result); James N. Farr et al., *Simplification of Flesch Reading Ease Formula*, 35 J. APPLIED PSYCH. 333 (1951) (revising the Flesch Reading Ease Formula to require counting only one-syllable words instead of syllables per word); George R. Klare et al., *Automation of the Flesch Reading Ease Formula, with Various Options*, 4 READING RES. Q. 550, 557-58 (1969) (counting vowels set off by consonants as syllables and using other algorithmic rules).

[112] See *infra* Appendix.

[113] Copies of the study in Adobe pdf format are available from the author.

[114] THE ART OF READABLE WRITING, *supra* note 30, at xi-xii (quoting THE ART OF PLAIN TALK, *supra* note 30, at xii).

Kincaid and his associates noted, "[a]ctually, readability formulas are only accurate to within one grade level . . . ."[115] Chall also recognized the limitations of readability tests:

> No readability formula is a complete and full measure of text difficulty. It measures only a limited number of the many characteristics that make text easy or hard to read and understand. An awareness of these limitations will lead to a wiser and more satisfactory use of readability measures. Hopefully, it will avoid a mechanical approach that can lead to disappointment.[116]

The limitations of these tests are not lost on everyone. The National Council of Teachers of English discourages the use of readability formulas in selecting materials for English language arts programs.[117] "Because readability formulas tend to be simplistic measures, such formulas should be used cautiously, if at all. Teachers' judgments about the difficulty of a work are more soundly based on complexity of plot, organization, abstractness of the language, familiarity of vocabulary, and clarity of syntax."[118]

Moreover, reading researchers have shifted their focus from formulas to empirical research.[119] According to one authority, "the nearly 30 year old Flesch-Kincaid scale and other readability formulas are considered antiquated by reading researchers."[120]

Nonetheless, legal professionals and researchers in other fields often retain faith in technological answers that are easy to access. For example, legislators acting in good faith have enacted statutes requiring consumer documents to be written at no higher than a particular grade level or even specifying the formula to apply in determining whether they meet that requirement. Yet, these legislators might have been unaware that applying different formulas to a document may yield different grade levels or that different versions of the same formula may also yield different results.

[115] KINCAID, *supra* note 3, at 20.

[116] CHALL & DALE, *supra* note 1, at 6.

[117] National Council of Teachers of English, *Guidelines for Selection of Materials in English Language Arts Programs*,
http://www.ncte.org/about/issues/censorship/five/116515.htm (last visited Sept. 12, 2007).

[118] *Id.*

[119] *See* Ancker, *supra* note 9, at 97.

[120] *Id.* "In browsing through my university's [Columbia University] psychology library, I found no texts on the psychology of reading that cited Kincaid's work or any other grade level scale" *Id.*

IV. CONCLUSION

This investigation offers four lessons. First, do not rely on technicians to follow the proper methodology in calculating a result on readability tests. Our primary example illustrates the point. Consider the cases[121] and studies[122] that relied on Microsoft to use the correct formula or at least divulge that it was using some alternative.

Second, conform to the proper test methodology. For example, in the study that generated the Flesch-Kincaid formula, the Kincaid researchers had subjects read sample texts averaging 170 words, but some of the samples were as short as 104 words.[123] It is unclear how accurate the test is for shorter passages. Yet researchers and courts sometimes employ much shorter passages,[124] which Word obligingly scores.

The Kincaid researchers also used a variety of short passages from a Navy manual. They did not attempt to score the entire manual. Scoring an entire document may overlook critical sections that are written at a very high grade level. Thus, in our initial example concerning the lease clause, different parts of the lease may even yield different scores.

Third, relying on a single source for critical information is a risky proposition. If the information is inaccurate, the results may be problematic. Yet, technology invites us to accept its single-source answers. In our case, determining readability requires the tedious process of counting sentences, words, and syllables and then multiplying, adding and subtracting. The computer, however, offers an automated method to compute these counts for us. How can we reject this technocentric offer? Not only is it difficult to refuse such an offer, but, unfortunately, we likely will also forgo confirming the accuracy of the software's algorithm.

[121] Given the mechanics of the test, it is safe to speculate that many of them do rely on Microsoft. *See, e.g.*, Conkling v. Keisling, 852 P.2d 183, 189 (Or. 1993) (Van Hoomissen, J., concurring) (noting that the official guide to help citizens understand a ballot proposition was written at the fourteenth grade level of education); Deras v. Roberts, 788 P.2d 987 (Or. 1990) (finding the proposed ballot title for an initiative measure failed to satisfy the statutory requirement for readability).

[122] *See, e.g.*, Cotugna et al., *supra* note 8; DuVivier, *supra* note 8, at 252 n.173; Nathaniel Good et al., *supra* note 8, at 343; Heilke et al., *supra* note 8; Kahn et al., *supra* note 8, at 131; Kimball & Kropf, *supra* note 8; Mailloux et al., *supra* note 8; Paasche-Orlow et al., *supra* note 8.

[123] *See* KINCAID, *supra* note 3, at 6, 26.

[124] *See, e.g.*, Paasche-Owen, *supra* note 8, at 723.

Fourth, just because a method is popular, it is not necessarily the best method. For example, common sense tells us that sometimes a sentence with few words and syllables can be difficult to read and a sentence with many words and syllables can be quite comprehensible. Consider Percy Bysse Shelley's *Ozymandias*.[125] It scores 6.6 on the Flesch-Kincaid scale (2.5 on the Word version). A.E. Houseman's *To an Athlete Dying Young*,[126] scores 7.49 (3.2 on the Word version). Grade school students and middle school students would be hard pressed to understand even the superficial meanings of these poems.[127]

A simple test prevailed because, in many cases, it is relatively accurate, and a popular commercial computer system has made it convenient to use. Whatever the test's limitations, however, those who offer it ought to present it in its correct form.

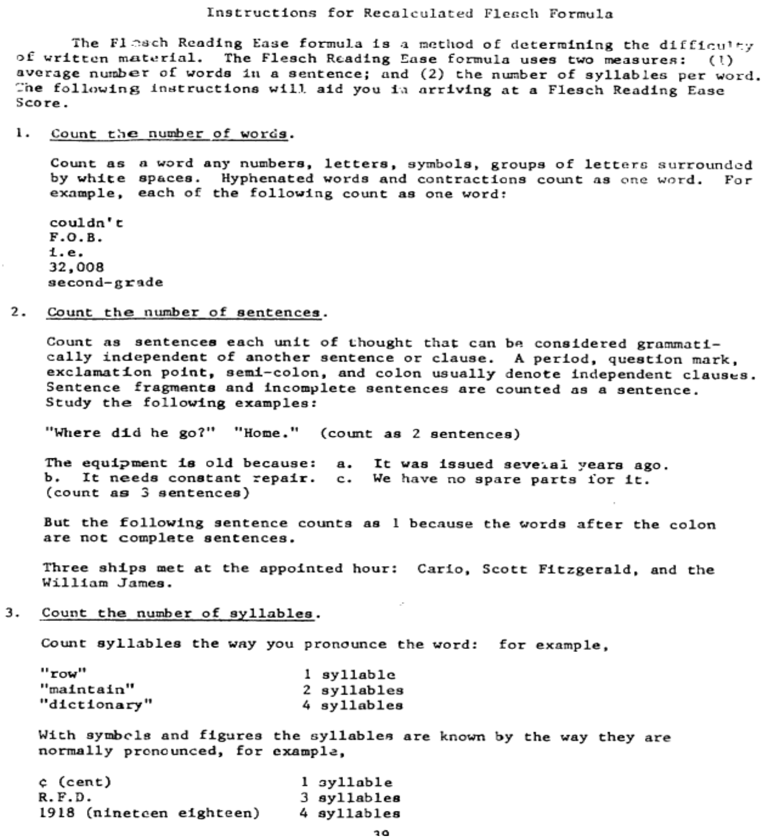
[125] Percy Bysse Shelley, *Ozymandias*, Poetry Out Loud, http://www.poetryoutloud.org/poems/poem.html?id=175903 (last visited Sept. 12, 2007).

[126] A.E. Houseman, *To an Athlete Dying Young*, Poetry Out Loud http://www.poetryoutloud.org/poems/poem.html?id=175749 (last visited Sept. 12, 2007).

[127] According to my calculations using the new Dale-Chall formula, "Ozymandias" ranks at the eleventh to twelfth grade level, and "To an Athlete Dying Young" ranks at the seventh to eighth grade level.

APPENDIX

Scoring Instructions for the Flesch-Kincaid Formula[*]



Instructions for Recalculated Flesch Formula

The Flesch Reading Ease formula is a method of determining the difficulty of written material. The Flesch Reading Ease formula uses two measures: (1) average number of words in a sentence; and (2) the number of syllables per word. The following instructions will aid you in arriving at a Flesch Reading Ease Score.

1. Count the number of words.

Count as a word any numbers, letters, symbols, groups of letters surrounded by white spaces. Hyphenated words and contractions count as one word. For example, each of the following count as one word:

couldn't
F.O.B.
i.e.
32,008
second-grade

2. Count the number of sentences.

Count as sentences each unit of thought that can be considered grammatically independent of another sentence or clause. A period, question mark, exclamation point, semi-colon, and colon usually denote independent clauses. Sentence fragments and incomplete sentences are counted as a sentence. Study the following examples:

"Where did he go?" "Home." (count as 2 sentences)

The equipment is old because: a. It was issued several years ago. b. It needs constant repair. c. We have no spare parts for it. (count as 3 sentences)

But the following sentence counts as 1 because the words after the colon are not complete sentences.

Three ships met at the appointed hour: Cario, Scott Fitzgerald, and the William James.

3. Count the number of syllables.

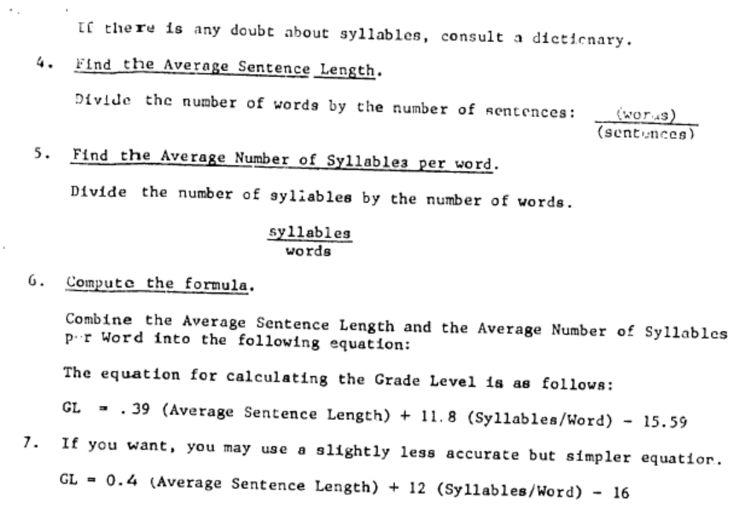
Count syllables the way you pronounce the word: for example,

"row"	1 syllable
"maintain"	2 syllables
"dictionary"	4 syllables

With symbols and figures the syllables are known by the way they are normally pronounced, for example,

¢ (cent)	1 syllable
R.F.D.	3 syllables
1918 (nineteen eighteen)	4 syllables

[*] KINCAID, *supra* note 3, at 39-40.



If there is any doubt about syllables, consult a dictionary.

4. Find the Average Sentence Length.

Divide the number of words by the number of sentences: $\frac{\text{(words)}}{\text{(sentences)}}$

5. Find the Average Number of Syllables per word.

Divide the number of syllables by the number of words.

$$\frac{\text{syllables}}{\text{words}}$$

6. Compute the formula.

Combine the Average Sentence Length and the Average Number of Syllables per Word into the following equation:

The equation for calculating the Grade Level is as follows:

GL = .39 (Average Sentence Length) + 11.8 (Syllables/Word) − 15.59

7. If you want, you may use a slightly less accurate but simpler equation.

GL = 0.4 (Average Sentence Length) + 12 (Syllables/Word) − 16